PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :
C12Q 1/68

A1

(11) International Publication Number: WO 98/13521

(43) International Publication Date: 2 April 1998 (02.04.98)

(21) International Application Number: PCT/EP97/05290

(22) International Filing Date: 26 September 1997 (26.09.97)

(30) Priority Data:
9620216.3 27 September 1996 (27.09.96) GB

(71) Applicant *(for all designated States except US)*: FONDAZIONE CENTRO SAN RAFFAELE DEL MONTE TABOR [IT/IT]; Via Olgettina, 60, I-20132 Milano (IT).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: CONSALEZ, Giangiacomo [IT/IT]; Via Olgettina, 60, I-20132 Milano (IT). FESCE, Riccardo [IT/IT]; Via Olgettina, 60, I-20132 Milano (IT).

(74) Agent: MINOJA, Fabrizio; Studio Consulenza Brevettuale, Via Rossini, 8, I-20122 Milano (IT).

(81) Designated States: JP, US, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: METHOD FOR THE DIFFERENTIAL SCREENING OF GENE EXPRESSION BY RANDOM PRIMED REVERSE TRANSCRIPTION-POLYMERASE CHAIN REACTION

(57) Abstract

The present invention concerns a method for the differential screening of gene expression by random primed Reverse Transcription-Polymerase Chain Reaction (RT-PCR) and a kit to be used for the performance of said method.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania
AM	Armenia
AT	Austria
AU	Australia
AZ	Azerbaijan
BA	Bosnia and Herzegovina
BB	Barbados
BE	Belgium
BF	Burkina Faso
BG	Bulgaria
BJ	Benin
BR	Brazil
BY	Belarus
CA	Canada
CF	Central African Republic
CG	Congo
CH	Switzerland
CI	Côte d'Ivoire
CM	Cameroon
CN	China
CU	Cuba
CZ	Czech Republic
DE	Germany
DK	Denmark
EE	Estonia
ES	Spain
FI	Finland
FR	France
GA	Gabon
GB	United Kingdom
GE	Georgia
GH	Ghana
GN	Guinea
GR	Greece
HU	Hungary
IE	Ireland
IL	Israel
IS	Iceland
IT	Italy
JP	Japan
KE	Kenya
KG	Kyrgyzstan
KP	Democratic People's Republic of Korea
KR	Republic of Korea
KZ	Kazakstan
LC	Saint Lucia
LI	Liechtenstein
LK	Sri Lanka
LR	Liberia
LS	Lesotho
LT	Lithuania
LU	Luxembourg
LV	Latvia
MC	Monaco
MD	Republic of Moldova
MG	Madagascar
MK	The former Yugoslav Republic of Macedonia
ML	Mali
MN	Mongolia
MR	Mauritania
MW	Malawi
MX	Mexico
NE	Niger
NL	Netherlands
NO	Norway
NZ	New Zealand
PL	Poland
PT	Portugal
RO	Romania
RU	Russian Federation
SD	Sudan
SE	Sweden
SG	Singapore
SI	Slovenia
SK	Slovakia
SN	Senegal
SZ	Swaziland
TD	Chad
TG	Togo
TJ	Tajikistan
TM	Turkmenistan
TR	Turkey
TT	Trinidad and Tobago
UA	Ukraine
UG	Uganda
US	United States of America
UZ	Uzbekistan
VN	Viet Nam
YU	Yugoslavia
ZW	Zimbabwe

METHOD FOR THE DIFFERENTIAL SCREENING OF GENE EXPRESSION BY RANDOM PRIMED REVERSE TRANSCRIPTION-POLYMERASE CHAIN REACTION

The present invention concerns a method for the differential screening of gene expression by random primed Reverse Transcription-Polymerase Chain Reaction (RT-PCR) and a kit to be used for the performance of said method.

The analysis of differential gene transcription focusses on molecular mechanisms involved in major biological processes, such as cell differentiation, cell division, embryonic development and neoplastic transformation. A multitude of techniques has become available in recent times to isolate differentially expressed genes. These techniques can be grouped in two classes: subtractive hybridization and differential screening. Hybridization-based differential screening and subtractive techniques are extensively covered elsewhere (1).

In 1992, Liang and Pardee first described a new, RT-PCR-based differential screening technique which they named Differential Display (DD) (2). In this technique, cDNAs are synthesized by means of anchored oligo-dT primers to select subsets within given mRNA populations. First strand cDNAs are subsequently PCR-amplified using the same downstream oligo-dT primer and an upstream random decamer. The complex PCR product is separated through a polyacrylamide gel and detected by autoradiography thanks to the incorporation in the PCR reaction of a radioactive dNTP. The technique aims at pinpointing bands corresponding to differentially expressed genes from a background of ubiquitous, constitutively expressed products. Various refinements of the above protocol have been published (3-5). An inherent limitation of the DD technique is the fact that products obtained from DD gels derive almost exclusively from noncoding regions of genes, making sequence analysis hardly informative. Furthermore, aside from the EST database, which mostly contains human sequence, only limited information is deposited into nucleotide sequence databases regarding 3' UTR regions. Consequently, a lengthy cDNA walk originating from the 3' end of a transcript can - frustratingly - result in the cloning of known coding sequence. Finally, problems can arise from the low sensitivity of the DD technique, which mostly identifies medium to high abundance transcripts, probably due to the use of low-complexity (oligo-dT) primers for PCR amplification.

Around the same period of time, a different RNA fingerprinting protocol was developed by other authors (6), to permit internally primed PCR amplification of oligo-dT-primed or random-primed cDNAs. In this protocol, named RAP-PCR, only arbitrary primers are used for the radioactive PCR amplification step. Preliminary data (GGC, unpublished observations) clearly indicate that this procedure leads to greater sensitivity, improved amplification and cloning efficiencies, and that a significant share of cDNAs lie in coding regions, making their sequence analysis considerably more informative and allowing some degree of prediction to be made as to their nature and possible function. However, unlike DD, arbitrarily primed RNA fingerprinting has not been systematized to maximize coverage of genes expressed in a given tissue or cell line within a discrete number of PCR amplifications. In other words, the main problem with RAP PCR lies within the unavailability of a rationally designed panel of primers permitting an exhaustive, nonredundant survey of gene expression in a given biological system.

The present invention solves the above problem by means of a computer-assisted search for RNA fingerprinting primers characterized by high amplification efficiencies and a marked, nonrandom affinity for coding regions. The collection of reagents generated according to the invention allow the use of internally primed, PCR-based RNA fingerprinting as a reasonably simple, exhaustive and systematic tool for the analysis of differential gene expression, and as a workable, advantageous alternative to differential display.

The method of the invention is characterized in that the PCR is carried out using a plurality of oligonucleotide primers the sequence of which has been determined by a method comprising the following steps:

a) generation of random primer sequences having a CG/AT ratio of 2:1, no stop codon, no more than three consecutive identical nucleotides and no palindromic 5' and 3' ends;

b) screening of the primer sequences generated in a) by simulating PCR reactions on non-redundant mammalian nucleotide sequence databank entries containing at least 1,000 bp of coding region and calculating for each primer sequence their:

(i) efficiency index, said efficiency index being defined as the ratio of the number of PCR products comprising coding sequences obtained using said primer sequence to the modal number of PCR products comprising coding sequences obtained for each of the whole set of tested primers generated in a); and

(ii) selectivity index, said selectivity index being defined as the ratio between the probabilities of yielding a PCR product comprising coding sequences or 3' untranslated regions; and

c) selecting some or all of the primer sequences screened in b) according to their efficiency index and selectivity index for use in PCR.

The invention also provides a kit for differential screening of gene expression in biological samples by means of random priming RT-PCT comprising:

a) a plurality of oligonucleotide primers selected according to the above described method;

b) reagents for the reverse transcription and amplification reactions;

c) optionally, protocols for the cloning of the products of differential screening.

The primers selected according to the criteria of the claimed method allow the detection of more than 80% of cDNAs containing significant portions of coding regions, compared with about only 10% of cloned products containing translated regions obtainable according to the prior art methods.

The present invention provides therefore a useful tool allowing easier recognition of new sequences as well as an easier comparison between known genes and the new genes cloned by the method of the invention.

Brief Description of the Figures

Figure 1. Histogram of the number of simulated PCR products (in CDS) per primer, tested on human nonredundant pseudo-cDNA database. Also shown is the expected distribution of number of products per primer, based on the probability of matching after randomly scrambling the sequences in the database (dashed line).

Figure 2. Scatter plot of the number of simulated CDS PCR products yielded by each primer when tested on the human (abscissa) or mouse (ordinate) pseudo-cDNA databases (454 primers).

Figure 3. Exhaustivity and redundance analysis on simulated PCR (96 most efficient primers tested on human nonredundant pseudo-cDNA database). Panel **A** shows the distribution of the number of simulated PCR products per transcript. Solid line: expected distribution, based on the probability of matching to the randomly scrambled sequences in the database. Dashed line: expected distribution for an increase in theoretical probability of matching by a factor equal to the ratio observed/expected mean number of products per transcript. Panel **B** shows the distribution of the number of different primers yielding simulated PCR products from each transcript. Dashed line: expected distribution after correcting the theoretical matching probability as in panel A.

Figure 4. Simulated PCR using pairs of dodecanucleotide primers degenerate at the 3'end nucleotide, on a nonredundant pseudo-cDNA database consisting of 2560 transcripts (5.4 million basepairs). Distribution of the number of different primer pairs yielding simulated PCR products from each transcript. Solid and dashed lines are the expected distributions before and after correcting the theoretical probability of matching by a factor equal to the ratio observed/expected mean number of products per transcript.

Figure 5. Correlation between the number of simulated PCR products and the number of bands in experimental gels. Simulations performed on mouse nonredundant pseudo-cDNA database, using 12-nt primers (8 C-G, 4 A-T). Thirteen primers were arbitrarily chosen among those yielding low, medium and high numbers of simulated PCR products; PCR experiments were performed as described in the Methods and the numbers of clearly discernible bands were recorded. The line illustrates the least-square linear regression on the data through the origin (vertical bars = S.D., r = 0.87).

Figure 6 shows sample gels obtained through computer-driven fingerprinting (RF) experiments. Uninduced Hep G2 cells and some line induced with reducing agents are compared. RNA extractions, RT reactions and PCRs are done in duplicate. U1, U2: uninduced; 11, 12: induced.

Detailed Description of the Invention

The random sequences described in a) of the above method can be generated easily using straight-forward computer algorithms.

Simulation of PCR reactions as in b) of the above method can be performed by, for example, searching both strands of the target sequence for a sequence complementary to the primer sequence, permitting varying degrees of mismatch, for example 3 mismatches. A PCR product is scored if a suitable match is found on both strands and the matching sequences are within a predetermined distance from each other, for example from 100 to 1,000 bp apart. Searches can be performed using using any of several commercially available software packages, such as FINDPATTERNS in the Wisconsin GCG package.

Non-redundant nucleotide sequence databases are used to provide target sequences for PCR simulations. Nucleotide sequence databases are easily accessible to skilled person, two of the largest and most well-known being the Genbank and EMBL databanks. From these databanks, a subset of sequences are selected. Only sequences containing at least 500 bp, preferably at least 1000 bp, of coding sequence are selected. Furthermore redundant sequences are eliminated. That is to say, for any given gene, often more than one entry occurs in the databank and it is desirable to select only one of the entries if they all have very similar sequences. One method of achieving this is to compare databank entries which have a common word in their sequence descriptions with a sequence comparison program, for example FASTA, and eliminate the shorter sequence if the two sequences have sequence identity above a percentage threshold, preferably greater than 95%. It is also desirable to eliminate intron sequences from genomic sequences to produce a contiguous cDNA sequence.

The oligonucleotide primers may be of any length and preferably comprise from at least 10 to 20 nucleotides, for example they may consist of 12, 15 or 18 nucleotides, more preferably 12 nucleotides. Primers may contain additional groups such as labelling groups, for example biotin, or radio-labelled substituents. Primers may be synthesised using standard methods known to those skilled in the art, for example using an automated oligonucleotide synthesiser.

In order to obtain more readable PCR-amplification gels, the efficiency index as defined above and used as one of the two criteria for selecting candidate primers should not be either too low (preferably > 2) or too high (preferably ≤ 10). The selectivity index as defined above and used as the other criterion for selecting candidate primers is preferably higher than 1, more preferably higher than 1.8, even more preferably higher than 2.

Some primers may produce PCR products containing both coding sequences and untranslated regions. However, the amount of coding sequence in some cases may be very small. Therefore, when determining efficiency and selectivity scores it may be preferable to only consider a primer as having yielded a product within a coding sequence if the amount of coding sequence within the product exceeds a predetermined percentage, for example 10, 30, 50 or 70%, or a predetermined length (e.g. 50, 100 or 200 nts).

The set of primers selected according to the method of the invention described above may be further selected from to produce a smaller set of primers. This can be accomplished by simply selecting primers with the highest selectivity indices. For example, if 500 primers are selected after steps a) to c), it would not be necessary to synthesise all 500 primers for use in PCR techniques. A skilled person may only select, for example, from 10 to 100 primers. Generally, a skilled person would select primers with the highest selectivity index and efficiency index except that they may discard any primers with sequences that are too similar to other selected primers, for example if they are greater than 80% identical overall (or greater than 60% identical in the last 8 nucleotides at the 3' end).

The kit described above will typically contain about from 10 to 200 primers, or pairs of primers degenerate at one position (e.g. the last nucleotide at the 3' end) preferably from 20 to 100 primers (or pairs), for example 30, 60 or 96 primers (or pairs), selected by the method of the invention.

The method of the invention is hereinafter described, by way of an example, in more detail with reference to specific databases and experimental conditions.

a) **Simulation of mRNA PCR in nucleotide databases**

PCR simulations were run on two nonredundant (nr) databases, obtained from a combination of human or mouse sequences deposited into the Genbank and EMBL nucleotide sequence databanks (accessed through the GCG Wisconsin package, version 8.1-UNIX, August 1995) (7), using one arbitrary 12-nt primer sequence at a time, thus assuming each primer to anneal in a degenerate fashion to the sense and antisense strand.

The reduced human and mouse databases were obtained by selecting human or mouse sequences containing at least 1000 bp of coding region (CDS). In order to decrease redundance, variable regions of immunoglobulins and T-cell receptors were eliminated, and all pairs of sequences sharing a word in their product descriptions were compared by the FASTA algorithm (8); the shorter one was eliminated when >95% identical to the other. Intronic regions were eliminated from genomic sequences, generating new transcribed sequence files containing uninterrupted cDNA.

Annealing of the primers was simulated by searching both strands for the sequence of each primer in the nr databases by means of the FINDPATTERNS program in the Wisconsin GCG package (7), permitting a maximum of 3 mismatches. All pairings with one or more mismatched base(s) among the last 4 (at the 3' end) were excluded as unsuitable to prime a polymerase chain reaction (PCR). A simulated PCR product was scored whenever a pairing occurred on the sense strand and, 100-1000 bp downstream, on the antisense strand.

For each primer, simulated PCR products were tagged with a CDS flag, if they contained a coding sequence portion, a UTR flag, if they contained a portion of 3' untranslated region. Each primer could be assigned an "efficiency" score (total number of simulated PCR products in the sequence database) and a "selectivity" score (ratio of the probabilities of yielding a PCR product comprising coding sequences or untranslated -3' regions). A crucial aspect in assessing the validity of the approach proposed here is to exclude the possibility that differences in "efficiency" observed among random primers are due to chance. Thus, the distribution of the number of products per primer, obtained through the simulation experiments, was compared to the one expected for random primers and random sequences, considering that (i) all primers were constituted by 8 G/C and 4 A/T, (ii) each sequence in the databank had a certain proportion of G/C and (iii) perfect match was required for 4 bases at the 3' end whereas up to 3 mismatches were allowed over the first 8 bases from the 5' end of each primer. The computation of the expected products number distribution is described in the following section.

b) **Computation of the expected distribution of PCR products per primer**

Let the databank, **D**, be a set of N sequences,

$$\mathbf{D} = \{S_s | s \in [1, N]\}.$$

Given that S_S is a sequence composed of a_S C/G nucleotides and b_S **A/T** nucleotides, the probability of a G or C nucleotide in the primer matching an arbitrary nucleotide in the sequence is

$$\frac{a_s}{2(a_s + b_s)},$$

and the corresponding probability for **A** or **T** is

$$\frac{b_s}{2(a_s + b_s)}.$$

In order to obtain hybridization a certain degree of matching must be obtained; here we arbitrarily decided that hybridization would occur for at least 9 matching bases out of 12, with no mismatches within the last 4 bases at the 3' end. Under these conditions, the probability of hybridization for a given template sequence and a given primer is a function of the fraction of C/G nucleotides in the sequence, $F_S = a_S/(a_S+b)$, the number of C/G in the first 8 bases of the primer, n_1, and the number of C/G in the last 4 bases at the 3' end, n_2.

For any specific alignment of the primer on the template S_S, we have:

$$\mathrm{p}[\mathit{base}(j) \text{ of the primer matches} \mid F_s] = \begin{cases} F_s/2 & \mathit{base}(j) \in \{C,G\} \\ (1-F_s)/2 & \mathit{base}(j) \in \{A,T\} \end{cases}$$

$$A_s = \mathrm{p}[\text{at least 5 out of the first 8 bases match} \mid F_s, N_1] =$$

$$= \sum_{j=1}^{N_1} \mathrm{p}[j \text{ matches out of } n_1\ (\mathrm{C} \vee \mathrm{G})] \cdot \mathrm{p}[\text{at least } 5 - j \text{ matches out of } 8 - n_1\ (\mathrm{A} \vee \mathrm{T})\ ] =$$

$$= \sum_{j=1}^{n_1} \left[C_{n_1}^{j} (F_s/2)^j (1 - F_s/2)^{n_1 - j} \cdot \sum_{k=5-j}^{8-n_1} C_{8-n_1}^{k} \left(\frac{1-F_s}{2}\right)^k \left(1 - \frac{1-F_s}{2}\right)^{8-n_1-k} \right]$$

$$B_s = \mathrm{p}[\text{all 4 bases at the 3' end match} \mid n_2] = \left(\frac{F_s}{2}\right)^{n_2} \cdot \left(\frac{1-F_s}{2}\right)^{4-n_2}$$

so that the probability of hybridization for any specific alignment of the primer on the template is $P_S = A_S \cdot B_S$. The average value of F_S was 0.53 (± 0.082) and in general P_S was about $1\text{-}2\cdot10^{-4}$, its value increasing for primers with increasing numbers of C/G nucleotides in the last 4 positions.

Assuming that PCR products of interest would have a length, L, comprized between $L_0 = 100$ and $L_1 = 1000$ BP, then the number of combinations of two acceptable positions on a template sequence S_S, of length M_S is:

$$C_s = \sum_{j=1}^{M_s-L_1}(L_1-L_0+1)+\sum_{j=M_s-L_1+1}^{M_s-L_0+1}(M_s-L_0-j+1)=(M_s-L_1)\cdot(L_1-L_0+1)+\sum_{k=0}^{L_1-L_0}k=$$

$$=(L_1-L_0+1)\cdot\left(M_s-\frac{(L_1-L_0)}{2}\right)=(M_s-450)\cdot450.5$$

This gives rise to a binomial distribution of the number of PCR products obtained from template sequence S_S of length M_S and a primer with given values of n_1 and n_2. Such distribution is defined by the binomial parameters $\mathbf{p} = P^2{}_S$ and $\mathbf{n} = C_S$; the corresponding probability density function (p.d.f.) is:

$$p_s(x)=C_{C_s}^{x}\cdot P_s^x\cdot(1-P_s)^{C_s-x}$$

Actually, to estimate the probability of obtaining simulated PCR products (neglecting the technical aspects connected to experimentally obtaining a PCR amplification product), the unwanted possibility of a further hybridization in between the two valid positions must be excluded. For a product of length L, this possibility has an approximate probability of

$$\left(1-(1-P_s)^{2L}\right)\approx 2L\cdot P_s,$$

and therefore about $900\cdot P_S$ for the average product length of 450 bp. For the usual magnitude of P_S this factor amounts to about 10^{-2} and can be neglected.

The distributions of PCR products from the N sequences in the database

$$\{p_s(x)\mid S_s\in\mathbf{D}\}$$

are expected to be independent. Therefore, the corresponding characteristic functions (ch.f.) can be computed

$$\{\varphi_s(u)\mid S_s\in\mathbf{D}\}$$

and the ch.f. for the whole databank will simply be:

$$\varphi_{\mathbf{D}}(u)=\exp\left(\sum_{S_s\in\mathbf{D}}\log[\varphi_s(u)]\right).$$

From $\varphi_{\mathbf{D}}(u)$ the expected p.d.f. of the number of PCR products from the whole databank, $p_D(x)$, is computed for each primer. Averaging over the set of primers yields the expected distribution of the number of PCR products per primer (P_1). Notice that $p_D(x)$ is necessarily equal for primers having the same values of n_1, n_2 and P_S. Thus, P_1 may be multimodal (up to 5 peaks for $n_2 = 0$ to 4).

The same procedure is used to compute the expected p.d.f. of the number of PCR products from each sequence, P_2 (in this case the ch.f. is computed by summing the logs of the single ch.f.'s over the set of primers for the same sequence).

A third distribution of interest is that of the number of "successful" primers per sequence (i.e. yielding at least one PCR product from the sequence), P_3. This is computed in the same way using the modified p.d.f.,

$$\pi_s, \text{such that:} \left\{ \begin{array}{l} \pi_s(0) = p_s(0) \\ \pi_s(1) = \sum\limits_{j=[1,\infty]} p_s(j) \end{array} \right\}.$$

Distribution P_1 is used to check whether the observed distribution of PCR products per primer significantly departs from the expectation: if a marked excess of particularly "good" and "poor" primers are found, this argues against a purely random distribution of nucleotides in the sequences of the databank.

Distributions P_2 and P_3 yield information on the exhaustivity of the approach, i.e. the capability of picking out as many different sequences as possible. In particular, the shape of the p.d.f. P_3 can be compared to the corresponding distribution, obtained by the simulation experiments, to check whether any bias is present towards a subpopulation of sequences (i.e. whether some sequences are significantly more subject to amplification than others).

This approach cannot be straightforwardly applied to the distributions obtained using sets of particularly efficient primers. These distributions are obviously shifted to the right, with respect to the p.d.f. P_3, which is computed based on a purely random nucleotide composition of the databank. The size of the shift is conveniently represented by the ratio of the mean values. If the shift simply reflects an increased hybridization probability with no bias towards sequence subpopulations, the shape of the curve will be easily reproduced by computing the logarithm of the characteristic function of the expected probability, multiplying it by the ratio of the means and computing the resulting probability distribution (this is performed by applying the direct and inverse fast Fourier transforms). A reasonable agreement with the observed distribution will argue against biases in favour of specific sequence subpopulations.

c) **RNA fingerprinting**

Reverse transcription is carried out using a (dT)16 primer on 1 mg total RNA extracted by the caesium chloride method (9). Radioactive PCR reactions, in duplicate, are performed from 2 µl of each RT reaction in 50 µl final volume with arbitrary 12-mers (final conc. 4 mM), using Perkin Elmer 1 x Amplitaq polymerase + $MgCl_2$ [1.5 mM]. PCR conditions are 3 minutes at 94°C, 2 minutes at 80°C at which

Taq polymerase is added (hot start), followed by 35 cycles of 40 secs. at 94°C, 1 min. at 50°C, 1 min. at 72°C, with a final elongation step of 5 mins at 72°C. 0.2 µl [α-^{32}P]dCTP are added to each reaction. Amplified products are separated on a 5% denaturing polyacrylamide gel and visualized by autoradiography. Differentially displayed bands are cut from the gel and electroeluted in dialysis bags as described (2). Bands are reamplified using the same 12-mer primers and cloned into a modified pBluescript II SK+ (Stratagene). Clones corresponding to differentially displayed bands are selected from the background of unrelated products.

Sequence analysis of cloned products - Data bank searches (Genbank, GenEmbl, SwissProt and PIR) are run through the BlastN and BlastX network servers (10). Additional sequence analysis and contig assembly is done using the GCG package.

d) Simulation of RNA fingerprinting PCR in human and murine nr nucleotide databases

In a first series of simulations (MAN12.8) 10,000 12-character strings were generated randomly, to represent dodecanucleotide primers, with the initial requirement that they contain 8 C or G and 4 A or T. Primers containing either stop codons (TAA, TAG, TGA) in the sense strand or ≥4 homonucleotide stretches (AAAA, CCCC etc.) were discarded (criteria a and b). Also discarded were primers with palindromic 5' and 3' ends (≥4 successive complementary bases (criterium c) or containing >5/8 bases at the 3' end identical to a previously accepted primer (criterium d). The above criteria were aimed at biasing the primers towards the CDS (a), at enhancing the efficiency of PCR experiments (b and c) and at reducing the chance of targeting the same sequence repeatedly (d). Finally, the last position (3' end) was partially degenerate, consisting of a W (A/T) or of an S (C/G).

A series of 1000 acceptable primers (according to criteria a, b and c) were challenged against the human nr database to measure their efficiency (total number of simulated PCR products) and selective affinity for coding portions of transcripts. Only primers yielding >100 simulated PCR products out of 2,085-sequences of our human nr database (6.04 Mb total DNA, 72.8% CDS) were included in the "good primers" list (158 primers); 497 primers were discarded because of their similarity to previously included primers (criterium d). The remaining 345 primers yielded < 100 simulated PCR products.

Figure 1 illustrates a histogram of the number of simulated PCR products obtained from each primer in this series (503 primers). Also illustrated is the probability density function (%) expected based on the probability of matching to the randomly scrambled database sequences (see Section b for details on the computation of this curve). It can be clearly seen that the observed distribution does not fit a random distribution of bases in the sequences, and that the extreme shoulders of the distribution curve are markedly overcrowded. This indicates a large excess of particularly poor and particularly efficient primers, and points to the possibility of selecting efficient PCR primers based on the present "simulated gene fishing" approach.

The simulation was repeated (MOUSE12.8) by testing the same series of primers on the mouse nr nucleotide database, containing 1041 sequences comprised of 2.95 Mb total nucleotide sequence, (72.5% CDS), in order to check whether species-specific features in sequence composition played a major role in determining the efficiency scores of different primers. In this case 193 "good" primers and 326 inefficient ones were obtained; 481 primers were not considered due to criterium d. The distribution of the numbers of simulated PCR products per primer was qualitatively similar to the one obtained in the human database (not shown).

Figure 2 is a scatter plot of the number of simulated PCR products obtained in the human (x-axis) vs. the mouse (y-axis) database, with the same primers (454). The two sets of values correlate very well (correl. coeff. r = 0.947), indicating that the unexpectedly high or low efficiencies of some primers did not arise from aberrations in the composition of the particular database used for the simulation experiments, but rather from intrinsic differences in efficiency among primers. In other words, it appears that some "genetic strings" (the base composition of some particular oligonucleotides) have particularly high or low probabilities of occurring in mammalian coding sequences.

One important aspect regards the exhaustivity of the present approach. The primers selected because of their high efficiency in yielding simulated PCR products may be directed towards subpopulations of genetic sequences. If this were the case, the number of sequences not picked out (sequences from which no PCR products are obtained) should be higher than expected on the basis of the average number of primers giving rise to products in each gene.

Figure 3 illustrates this point; in particular, it shows the distributions of the numbers of PCR products generated by each primer in the human database (A), and the distribution of the numbers of different primers yielding at least one PCR product from each transcript (B). Data are relative to sequences containing at least 1000 bp of coding region. The dashed line in B represents the expected distribution of numbers of primers (among the 96 selected here) picking out each sequence, given the average value of such distribution (see Section b). The observed distribution reasonably agrees with this expectation; in particular, the percentage of sequences not picked up by any primer is reasonably well predicted. These graphs suggest that the sets of "efficient" primers proposed here are capable of targeting the entire population of sequences deposited into databanks.

In order to further check for possible biases in the procedure, the same kind of simulation was performed using 12-nucleotide primers composed of 6 A/T and 6 C/G (MAN12.6, MOUSE12.6). Again, the primers displayed a wider range of efficiencies than expected, and again the distribution of the numbers of primers yielding products from each gene was in reasonable in agreement with the predictions of a random-interaction model. As reported below, in actual experiments stringency conditions had to be very relaxed in order to obtain a reasonable number of PCR products using 12-nucleotide primers. Better results were obtained by using degenerate pairs of primers. As this approach might as well increase the exhaustivity of mRNA fingerprinting, we repeated the simulation using 12-nucleotide, 8-C/G primers, containing a partially degenerate base (W or S) at their 3' ends. This simulation was run on an updated database, and the results were similar to those reported above. Again, "poor" and "very good" primers are present in large excess with respect to the expectation for a random-sequence database. A set of 96 "best" primers was selected as those displaying an efficiency index comprised between 2 and 15 and a selectivity index above 1.4. Figure 4 shows that the best 96 primer pairs yield a distribution of number of primers spotting each sequence which is shifted to the right with respect to the expectation (solid line), suggesting that this subset of primers is particularly efficient; the distribution is well fit when the expectation is corrected for the mean efficiency of this particular set of primers (dashed line) and the number of sequences yielding no simulated PCR products.

d) **Experimental assessment of the primer panel**

As a result of the elaboration described above, a panel of 120 optimal degenerate primer sequences (Table 1) was generated. All these primer pairs yielded E.I. comprised between 2 and 20 and S.I. > 1.4. Thirteen primers (some belonging to this panel and some not) were synthesized and tested at the bench, to assess the correspondence between theoretical predictions and experimental results. As templates, we utilized oligo-dT-primed cDNAs obtained from various mouse and human cell lines. The results are shown in Figure 5.

Exhaustivity - So far, all primers have displayed amplification efficiencies coherent with those expected based on our simulation. Ten degenerate primers tested so far on a HepG2 cell line total cDNA, have produced patterns containing 80 to 167 bands (mean: 114.4 bands/gel) (**Figure 6**). NNN oligos predicted to amplify inefficiently in mammalian cDNA were also tested, and two gave very poor banding patterns (fewer than 30 bands). Extrapolating these preliminary data, it may be inferred that a differential screening study employing the best 120 primers should permit a survey of 10982 bands. An estimate of the coverage provided by this figure is dependent upon the degree of redundance and the complexity of the gene pool in each given tissue or cell line. At this stage, an experimental assessment of redundance is of limited significance, due to the small number of products analyzed so far. To date, however, the ten cDNAs cloned with this new set of primers came from ten distinct genes.

Selectivity for coding regions - cDNAs cloned from these experiments contained ORFs throughout their lengths in 9 out of 10 cases. Of the nine ORFs, five corresponded to known coding regions (known genes or orthologs of known genes). Of the remaining 5 ORFs, four were rated "excellent" by the GRAIL program [11, 12], predicting coding regions within them; finally, although the tenth clone was n recognized as 'coding' by GRAIL, it also contained a 160 bp ORF at one end.

Accuracy - As far as the correspondence between RNA levels predicted through our intenally primed RNA fingerprinting protocol and actual transcript levels, out of 28 genes cloned with this method with the old (nondegenerate) set and new (degenerate) set of primers, 25 have been found differentially expressed as expected. RNA studies have been performed using cloned "differentially displayed" bands as probes or sources of primers for Northern analysis, RNase protection assay, quantitative RT-PCR).

Sensitivity - Besides assessing the above issues, we tried to determine whether our internal primers exhibit a preference for highly abundant transcripts (sensitivity) in at least two cases, cDNA clones generated by our method produced no hybridization signal by northern analysis of 30 mg total RNA. In two cases, cDNAs found by RT-PCR in a given tissue have required plating of over 10^6 pfu to permit isolation of the corresponding gene by hybridization-based screening of the appropriate library.

TABLE 1

No.	sequence	S.I.	E.I.	No.	sequence	S.I.	E.I.	No.	sequence	S.I.	E.I.
942	ACGCCATCGACC/G	6.39	1.84	115	GGCGCCTACTTC/G	2.23	5.09	114	CTGGTCACGTGC/G	1.64	3.88
688	AAGCTGCTCGCG/C	4.99	2.07	674	CGGCTTTCGTGC/G	2.16	2.22	224	CCCGTCTACCAC/G	1.63	12.37
522	GGCACATTGCGG/C	4.5	1.96	20	ATGCCCAAGCGC/G	2.13	2.11	313	GCCCTTCGTCAG/C	1.63	5.53
567	CCAGATGCCCGA/T	4.44	2.06	470	CAGCAAGTCGGC/G	2.09	2.81	149	CACAACCGGCAG/C	1.63	10.82
328	GCAGCATCCGGA/T	3.86	2.75	176	AAGCCGACCTCC/G	2.08	5.79	453	GTGCCACGCATC/G	1.62	2.6
60	CGATCACAGCGG/C	3.63	2.28	156	GCGCATCCAGCT/A	2.08	14.82	119	GCTCGATCAGGC/G	1.6	6.3
95	TCGATGCCGCTG/C	3.35	8.03	109	GGCATTCCGCAC/G	2.04	4.32	542	AGGACGCGCATC/G	1.6	3.12
417	ATGGCAACGGCG/C	3.32	5.77	320	GCACCGACTGGT/A	2.02	9.16	580	TGGCCAACTGCG/C	1.59	3.93
2	TCTGGGAACCGG/C	3.19	2.65	137	ACCAAGCCCACC/G	1.99	11.2	185	TCGTCACCAGCC/G	1.59	6.81
187	TGCTGCAGGACC/G	3.17	7.61	138	CCTGTCCGTCCT/A	1.93	18.49	463	GTGGACGGTGCA/T	1.59	6.2
88	CGTGGGCAACCT/A	3.13	6.9	230	GGTCCCAATGGG/C	1.93	10.68	132	TCGTGGCTGCAG/C	1.58	12.19
814	GAGCTTCACCGC/G	3.13	2.32	357	GTCCTGCGGGTT/A	1.92	1.91	290	GCCTCCTGGAGT/A	1.56	9
780	GAGGCGACGATC/G	3.1	3.92	423	AGCGGAGCATCC/G	1.92	1.91	56	TGGCTGGGATGG/C	1.54	8.4
402	TCGTCGACGGTG/C	3.08	1.89	283	AGCTCTGCGAGC/G	1.91	7.73	332	GCGGATGCGGAA/T	1.53	4.02
282	ACAGGCGATGCC/G	3.08	3.06	23	CCGCATGTCCAC/G	1.9	19.46	163	ACGTGCCCAGCA/T	1.53	11.55
648	AGAGCAGGGCGA/T	2.93	2.01	560	CGCCCTGGAACT/A	1.9	3.07	51	TGCCGACTCTGC/G	1.51	15.7
91	CTTCTCCCGGTC/G	2.8	4.82	868	GTCGCCGGCAAT/A	1.89	1.83	245	CTGCAGGGACCT/A	1.5	9.93
685	TGTGGAGCCGGT/A	2.79	2.38	755	GCTGCCGCCAAT/A	1.88	3.52	615	GAACACAGCCGG/C	1.5	2.68
151	CGGCACATCTCC/G	2.68	4.95	219	CGCTGTCGAGGA/T	1.87	5.99	334	TCGCCAGGATGC/G	1.49	7.9
333	GGCCGCATTGGA/T	2.66	6.48	595	AGGGCTTTCGGC/G	1.87	2.34	41	GGAGGAGAGCCA/T	1.49	13.01
97	GCAGAAGCCGTG/C	2.64	2.6	130	GGTGCTCAGCAG/C	1.85	15.4	370	GGCTGCACTTGC/G	1.49	6.66
29	CGGTCATGGTCG/C	2.58	2.26	96	CCTTGGAAGCCC/G	1.82	8.35	33	GGGTATGTGGCC/G	1.48	6.24
309	GGCCGAAGACCA/T	2.57	5.91	507	GCGGTCGAAGAC/G	1.82	4.33	31	CGCCTCATTGCG/C	1.48	2.55
850	CCAGCACTTCGC/G	2.54	2.17	324	TCTGCCGGGTCT/A	1.82	6.8	42	CGCCAATTGCCG/C	1.47	2.47
468	AGCCATTCGGGC/G	2.53	2.86	192	AGCCGGAGGATG/C	1.81	7.5	257	CCCGTCTCACCA/T	1.46	6.44
80	GTGTTGGTGCCC/G	2.47	7.53	745	AGAGTGCGCTCG/C	1.79	2.77	27	GTGGAGAGCTGC/G	1.46	19.58
180	TGGACGTTGGCC/G	2.42	8.05	112	CCTGCCGGAAGA/T	1.79	6.94	140	GCTCCAGTGGCA/T	1.46	16.79
24	GGAGAAGCTGCC/G	2.39	10.91	952	CAGGCCACATCG/C	1.76	2.74	66	TGCGGCGAGAAC/G	1.46	4.45
824	AGGCGGACATCG/C	2.35	2.82	791	TCCTGCGGATCG/C	1.76	2.98	842	TGGGTTCACCCG/C	1.46	2.37
195	AGCAGCTCGTGG/C	2.35	8.91	164	CAGGTCACGGAG/C	1.74	7.74	298	GCCTGTGGCATG/C	1.45	6.55
52	CAATACGGGCCC/G	2.35	1.97	918	CCGGAAAGCACG/C	1.74	2.51	18	CCTGCACCGAGA/T	1.45	3.45
511	ACAAGGGCACGG/C	2.35	1.87	108	CTTGGTCGTGCC/G	1.72	2.71	501	CACCGTACGCTG/C	1.45	2.79
280	GCCGGGAACTTC/G	2.34	5.9	292	CGTGCAGTTCCC/G	1.71	3.55	12	AAACCGGCGCCA/T	1.45	2.56
103	ATCCTGCACCGC/G	2.33	3.28	372	TCTCCGCGGTCA/T	1.71	4.27	379	GGCTGGCATTGC/G	1.43	6.99
529	AGGTACCCGTGC/G	2.31	4.98	799	AGACCGTGGACG/C	1.7	2.33	223	GGGCAGCCTTCA/T	1.43	7.76
445	TGTTGTGGCGGC/G	2.3	3.01	366	GAGGAACCGGAC/G	1.68	10.69	131	GGATCAGCAGGG/C	1.43	19.86
701	CGGCTATCGGCT/A	2.28	2.2	286	GACCACCGTGTG/C	1.68	6.21	891	CGCCTCCAGCTA/T	1.41	4.62
302	GCGGACCTCATG/C	2.27	7.68	355	CCAGCGTGTTCC/G	1.68	2.12	196	CGCCAGCATCAC/G	1.41	10.3
212	CTCTCCGATGCC/G	2.26	3.7	910	CACCTACCGAGC/G	1.66	2.57	139	AGGTGGGTGGTG/C	1.4	8.82
237	CCGGGTCCTCAT/A	2.24	6.79	418	CCAACGAGTCCC/G	1.65	2.45	410	CCGAGACATGCC/G	1.4	1.95

REFERENCES

1. Ausubel, F.M., Brent, R., Kingstone, R.E., Moore, D.D., Smith, J.A. and Struhl, K. (1995) Current Protocols in Molecular Biology.

2. Liang, P. and Pardee, A.B. (1992) Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. Science, 257, 967-971.

3. Bauer, D., Muller, H., Reich, J., Riedel, H., Ahrenkiel, V., Warthoe, P. and Strauss, M. (1993) Identification of differentially expressed mRNA species by an improved display technique (DDRT-PCR). Nucleic Acids Research, 21, 4272-4280.

4. Liang, P. (1993) Distribution and cloning of eukaryotic mRNAs by means of differential display: refinements and optimization. Nucleic Acids Research, 21, 3269-75.

5. Liang, P. (1994) Differential display using one-base anchored oligo-dT primers. Nucleic Acids Research, 22, 5763-5764.

6. Welsh, J., Chada, K., Dalal, S.S., Cheng, R., Ralph, D. and McClelland, M. (1992) Arbitrarily primed PCR fingerprinting of RNA. Nucleic Acids Research, 20, 4965-4970.

7. Devereux, J., Haeberli, P. and Smithies, O. (1984) A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Research, 12, 387-395.

8. Pearson, W.R. (1994) Using the FASTA program to search protein and DNA sequence databases. Methods in Molecular Biology, 25, 365-389.

9. Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) Molecular Cloning: A Laboratory Manual. Cold Spring Harbor.

10. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. J. Mol. Biol., 215, 403-410.

11. Roberts, L. (1991), Science, 254, 805.

12. Lopez R., Larsen F., Pryd Z. (1994), Genomics, 24, 133-136.

CLAIMS

1. A method for the differential screening of gene expression in biological samples by means of random priming RT-PCR, characterized in that the PCR is carried out using a plurality of oligonucleotide primers the sequence of which has been determined by a method comprising the following steps:

a) generation of random primer sequences having a CG/AT ratio of 2:1, no stop codon, no more than three consecutive identical nucleotides and no palindromic 5' and 3' ends;

b) screening of the primer sequences generated in a) by simulating PCR reactions on non-redundant mammalian nucleotide sequence databank entries containing at least 1,000 bp of coding region and calculating for each primer sequence their:

(i) efficiency index, said efficiency index being defined as the ratio of the number of PCR products comprising coding sequences obtained using said primer sequence to the modal number of PCR products comprising coding sequences obtained for each of the whole set of tested primers generated in a); and

(ii) selectivity index, said selectivity index being defined as the ratio between the probabilities of yielding a PCR product comprising coding sequences or 3' untranslated regions; and

c) selecting some or all of the primer sequences screened in b) according to their efficiency index and selectivity index for use in PCR.

2. A method according to claim 1 wherein the oligonucleotide primers consists of 12 nucleotides.

3. A method according to claim 2 wherein the oligonucleotides primers consists of 8 C or G and 4 A or T.

4. A method according to any one of the previous claims wherein each oligonucleotide primer differs from other primers in at least 5 out of 8 bases at the 3' end.

5. A method according to any one of the previous claims wherein the simulated PCR reaction is carried out on non-redundant human or mouse data banks from which variable regions of immunoglobulins and T-cell receptors as well as intronic regions are eliminated.

6. A method according to any one of the previous claims wherein the oligonucleotide primers have an efficiency index between 2 and 10, and a selectivity index higher than 1.

7. A method according to any one of previous claims, wherein the oligonucleotide primers are partially degenerate at the last position of the 3'end.

8. A kit for differential screening of gene expression in biological samples by means of random priming RT-PCR comprising:

a) a plurality of oligonucleotide primers selected according to the method of claim 1;

b) reagents for the reverse transcription and amplification reactions;

c) optionally, protocols for the cloning of the products of differential screening.

9. A kit according to claim 8 comprising the oligonucleotide primers of Table 1.

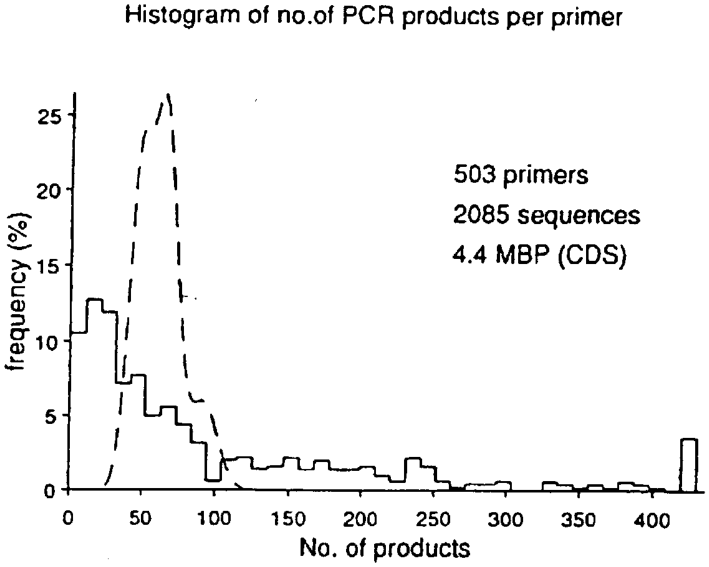

Histogram of no.of PCR products per primer
503 primers
2085 sequences
4.4 MBP (CDS)
frequency (%)
25
20
15
10
5
0
0
50
100
150
200
250
300
350
400
No. of products

FIGURE 1

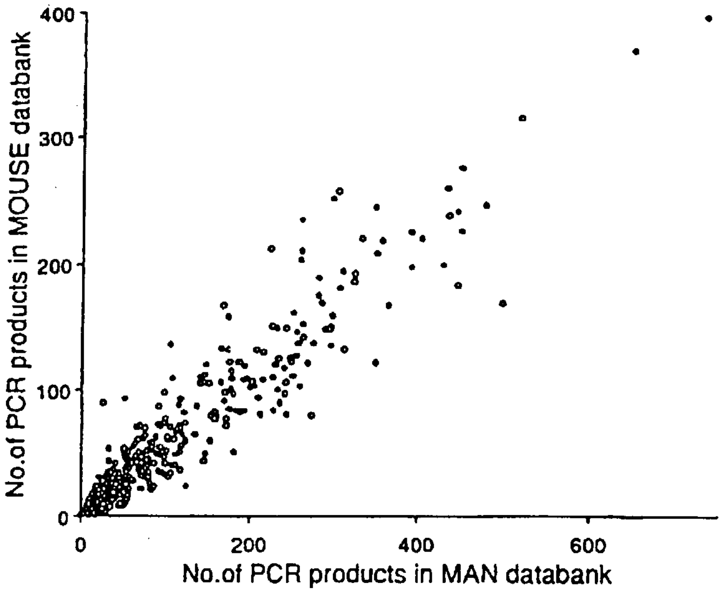

400
300
200
100
0
No.of PCR products in MOUSE databank
0
200
400
600
No.of PCR products in MAN databank

FIGURE 2

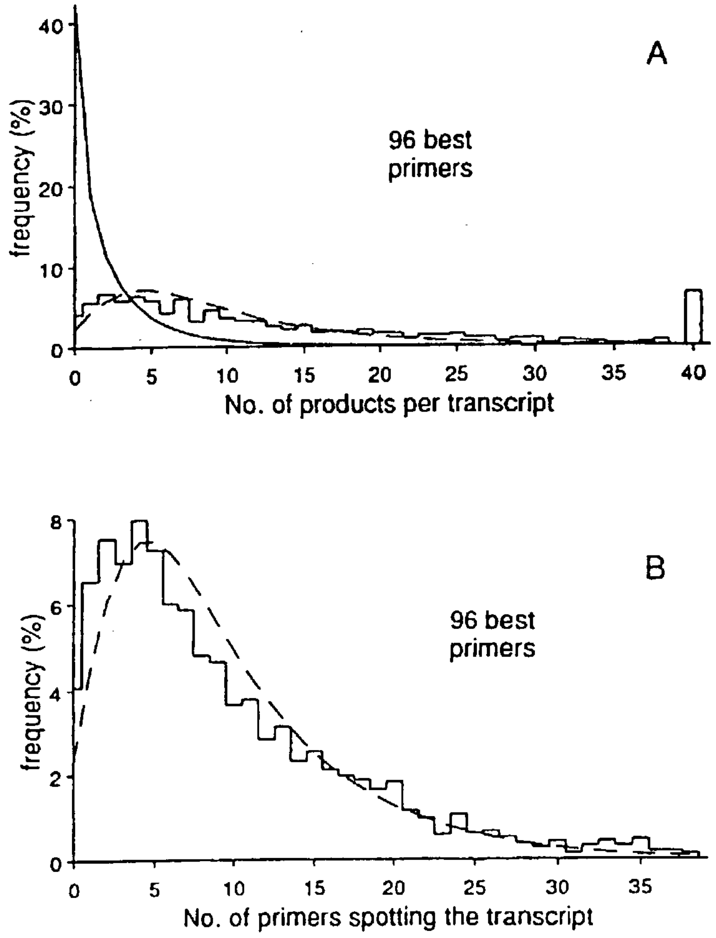

A
96 best primers
frequency (%)
0
10
20
30
40
0
5
10
15
20
25
30
35
40
No. of products per transcript
B
96 best primers
frequency (%)
0
2
4
6
8
0
5
10
15
20
25
30
35
No. of primers spotting the transcript

FIGURE 3

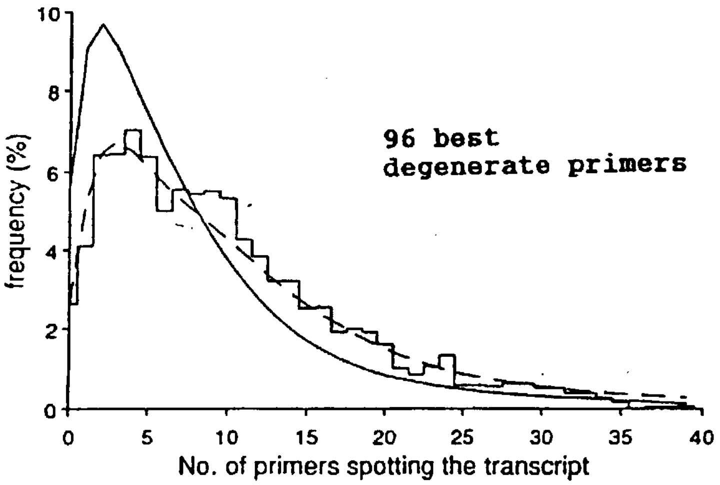

96 best
degenerate primers
frequency (%)
0
2
4
6
8
10
0
5
10
15
20
25
30
35
40
No. of primers spotting the transcript

FIGURE 4

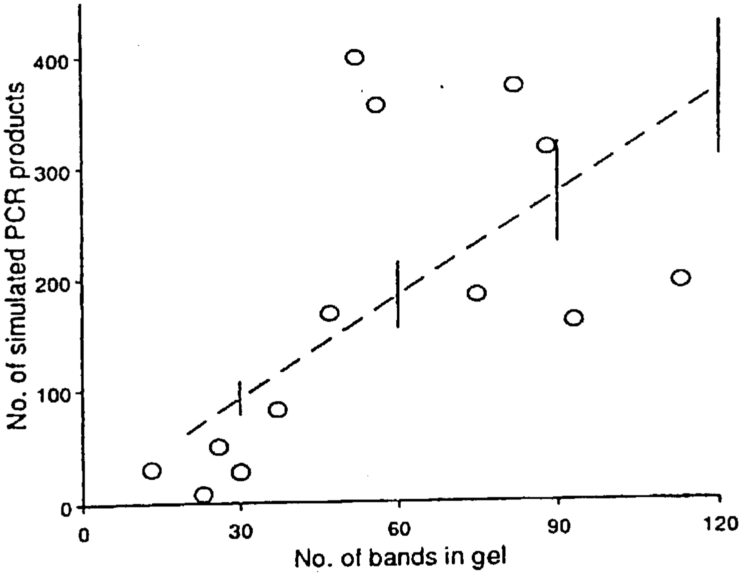
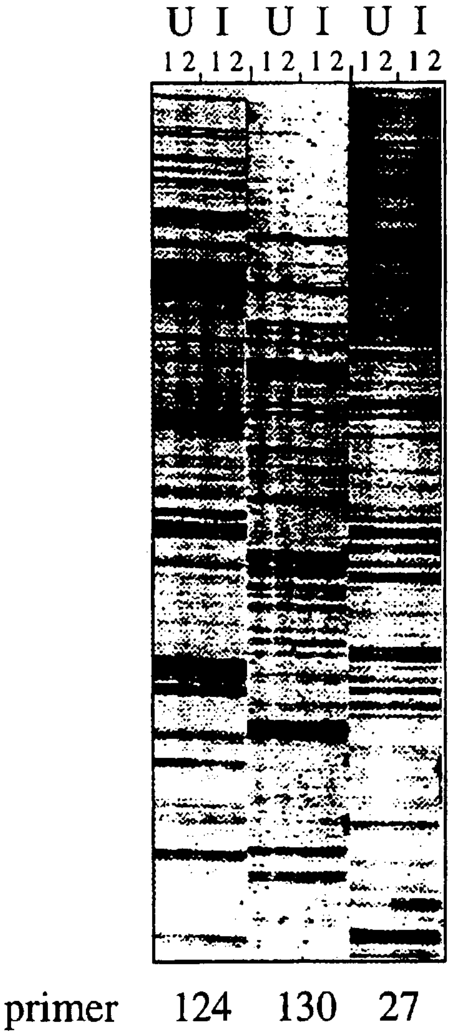

400
300
200
100
0
No. of simulated PCR products
0
30
60
90
120
No. of bands in gel

FIGURE 5



124 5'-AGCTTCGCCAGG
130 5'-GGTGCTCAGCAG
27 5'-GTGGAGAGCTGC

FIGURE 6

INTERNATIONAL SEARCH REPORT

Interr nal Application No
PCT/EP 97/05290

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 C12Q1/68

According to International Patent Classification(IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 6 C12Q

Documentation searched other than minimumdocumentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 95 33760 A (BRIGHAM & WOMENS HOSPITAL) 14 December 1995 see the whole document ---	1-9
X	LINSKENS M H K ET AL: "CATALOGING ALTERED GENE EXPRESSION IN YOUNG AND SENESCENT CELLS USING ENHANCED DIFFERENTIAL DISPLAY" NUCLEIC ACIDS RESEARCH, vol. 23, no. 16, 1995, pages 3244-3251, XP002047039 See page 3248, paragraph 2 see the whole document --- -/--	1-9

☒ Further documents are listed in the continuation of box C.

☒ Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publicationdate of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of theinternational search 6 March 1998	Date of mailing of the international search report 13/03/1998
Name and mailing address of the ISA European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Authorized officer Hagenmaier, S

INTERNATIONAL SEARCH REPORT

Inter nal Application No
PCT/EP 97/05290

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
X	BAUER ET AL.: "IDENTIFICATION OF DIFFERENTIALLY EXPRESSED mRNA SPECIES BY AN IMPROVED DISPLAY TECHNIQUE (DDRT-PCR)" NUCLEIC ACIDS RESEARCH, vol. 21, no. 18, 1993, pages 4272-4280, XP000394394 cited in the application see the whole document ---	1-9
A	AYALA M ET AL: "NEW PRIMER STRATEGY IMPROVES PRECISION OF DIFFERENTIAL DISPLAY" BIOTECHNIQUES, vol. 18, no. 5, 1995, pages 842-850, XP002033062 see the whole document ---	1-9
A	WELSH J ET AL: "ARBITRARILY PRIMED PCR FINGERPRINTING OF RNA" NUCLEIC ACIDS RESEARCH, vol. 20, no. 19, 11 October 1992, pages 4965-4970, XP000508271 cited in the application see the whole document -----	1-9

1

INTERNATIONAL SEARCH REPORT

Information on patent family members

Interr nal Application No
PCT/EP 97/05290

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 9533760 A	14-12-95	AU 2662395 A	04-01-96